DOCUMENT RESUME

ED 319 800 TM 015 096

AUTHOR Yen, Wendy M.; Candell, Gregory L.

TITLE Increasing Score Reliability with Item-Pattern Scoring: An Empirical Study in Several Score Metrics.

PUB DATE Apr 90

NOTE 24p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 17-19, 1990).

PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Elementary Education; *Elementary School Students; Item Response Theory; Scores; *Scoring; Test Items; *Test Reliability

IDENTIFIERS *Item Pattern Scoring; Number Right Scoring; *Three Parameter Model

ABSTRACT

Reliabilities are compared for two types of test score data: number correct, and item response patterns. Item-pattern scoring using three-parameter item response theory takes into account how many and which items a student answers correctly. This procedure theoretically results in greater reliability than does number-correct scoring. Empirical reliabilities of scores based on the two types of data were compared within each of three score metrics: (1) scale score; (2) number correct; and (3) grade equivalent. The reliabilities obtained were based on at least 900 students in grades 4 through 8 for five content areas (reading comprehension, language expression, spelling, mathematics concepts and applications, and science) at 10 testing times for 50 replications in each score metric. Item-pattern data produced more reliable scores than did number-correct data in 49 replications in the scale score metric, 50 in the number correct metric, and 49 in the grade equivalent metric. The increases in reliability produced by item-patterning scoring were equivalent, on the average, to that expected by a 15% to 20% increase in test length. Eight data tables and three figures are included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED319800

Increasing Score Reliability with Item-Pattern Scoring:

An Empirical Study in Several Score Metrics

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
WENDY YEN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Wendy M. Yen

Gregory L. Candell

CTB Macmillan/McGraw-Hill

BEST COPY AVAILABLE

Presented at the annual meeting of the National Council on Measurement in Education, Boston, April 1990.

TM015096

Abstract

Reliabilities of test scores can be influenced by two important factors: 1) the type of performance data used in the scoring process and 2) the metric in which the scores are expressed. This study compares reliabilities for two types of data: number correct and item response patterns. Item-pattern scoring using three-parameter IRT takes into account not only how many items a student gets correct, but also which items these are. This procedure theoretically results in greater reliability than number-correct scoring.

Empirical reliabilities of scores based on the two types of data are compared within each of three score metrics: scale score, number correct, and grade equivalent. These reliabilities were obtained based on at least 900 students for five content areas (Reading Comprehension, Language Expression, Spelling, Mathematics Concepts and Applications, and Science) at ten testing times (Fall and Spring of grades 4 through 8), for a total of 50 replications in each score metric.

Item-pattern data produced more reliable scores than number-correct data in 49 of 50 replications in the scale score metric, 50 of 50 replications in the number correct metric, and 49 of 50 replications in the grade equivalent metric. The increases in reliability produced by item-pattern scoring were equivalent, on the average, to that expected by a 15 to 20 percent increase in test length.

Increasing Score Reliability with Item-Pattern Scoring: An Empirical Study in Several Score Metrics

Two aspects of test scores have an important impact on reliability: the type of performance data that is used and the score metric. This study examines their separate effects on score reliability.

Performance Data

Scoring methods can differ in terms of the test performance data they use. Examples are number-correct scoring and Item Response Theory (IRT) item-pattern scoring.[1] Consider Table 1, which contains an example of number-correct scoring. Each number-correct (NC) score corresponds to a scale score, which in turn corresponds to a particular national percentile, grade equivalent (GE), and normal curve equivalent (NCE). Students who have the same number-correct score all receive the same scale score, percentile, etc.

Insert Table 1 about here

Three-parameter IRT item-pattern scoring uses more information about a student's performance than just the number-correct score. In item-pattern scoring, each student's scale score is that which has the highest probability of having produced that student's pattern of correct and incorrect answers. This scale score is determined by a computer algorithm that considers the statistical characteristics of the particular items the student got correct and incorrect.

Item-pattern scoring greatly expands the number of possible scale scores, adding rows to the scoring table, as in Table 2. For this example, only two of the hundreds of possible additional scale scores are shown. The table is entered with the student's item-pattern scale score, and the student's percentile, GE, and NCE are determined from that row. The relationship between the scale score and the derived (normative) scores is the same for item-pattern and number-correct scoring; for example, a scale score of 687 corresponds to a percentile of 33 for both scoring methods. In fact, the number-correct table is subsumed by the item-pattern table. What differs between number-correct and item-pattern scoring is how the student's scale score is obtained and the number of distinct scale scores possible.

Also shown in Table 2 is a number-correct score expected for each scale score ($\hat{NC}$), which unlike the observed number-correct score, can take on fractional values. $\hat{NC}$ is discussed more fully in the Method section.

Insert Table 2 about here

Using different performance data can provide different levels of information about student performance and can rank order students differently. For example, consider three students, all with observed number-correct (NC) scores of 12. They would all have a number-correct scale score of 687 (see Table 1), but based on item-pattern scoring the first may have a scale score of 680, the second a 687, and

the third a 693. When two- or three-parameter IRT is used and items differ in terms of their discriminations and/or have non-zero guessing levels (i.e., lower asymptotes), optimal item-pattern scoring can be used that will theoretically produce more test information, less measurement error, and greater reliability than number-correct scoring (Lord, 1980, Chapter 5). This theoretical advantage for item-pattern scoring (as measured by squared standard errors of measurement) applies to all metrics that rank order students the same way (Lord, 1980, Chapter 6).

Score Metrics

Examples of different score metrics used in educational measurement include number correct (NC), grade equivalent (GE), normal curve equivalent (NCE), percentile, and scale score (SS). Conversions from one score metric to another are done using tables, as in the examples in Tables 1 and 2. For a particular type of data, going from one metric to another does not change the rank order of students.[2] That is, within the number-correct scoring method, the student stays in the same row of the table for all metrics. Similarly, within the item-pattern scoring method, the student stays in the same row for all metrics. The different metrics are used in education, not because they rank order students differently, but because they have different meaning or usefulness for different applications.

Introductory measurement courses explain that the scale score, GE, number correct, and other metrics are not linearly related to each other. Some metrics are naturally bounded or limited. For example, grade equivalents higher than 12.9 typically are not reported in achievement testing. A student's number-correct score cannot be greater than the number of items in the test, regardless of how able that student is. The least bounded metrics, NCEs and scale scores, tend to be the most linearly related. Figure 1 shows examples of the relationships between the score metrics presented in Table 1. The relationships based on item-pattern scoring would look the same, except there would be more points "connecting the dots" in Figure 1.

Insert Figure 1 about here

Reliabilities

Reliabilities are commonly calculated using Pearson correlations of scores on alternate test forms. Pearson correlations are not affected by linear transformations of scores, such as multiplying every students' score by a constant. However, Pearson correlations can be affected by nonlinear transformations. Thus, different score metrics can be expected to produce different Pearson correlations, even when these metrics rank order the students in the same way and have the same rank order correlations. The use of different types of data can also affect the rank order of students, which can produce different Pearson correlations.

Because Pearson correlations can be influenced by both the score metric and the type of data, comparisons of test reliabilities across scores can confound differences in measurement accuracy (i.e., information) with differences in score metrics. Such confounding occurred in a study by Qualls-Payne and Hoover (1987), in which they compared Pearson correlations in the number correct metric based on

number-correct scoring with Pearson correlations in the scale score metric based on item-pattern scoring.

This study examines empirical reliabilities across metrics and types of data and separates their effects.

Method

Students and Tests

Test results were analyzed for a random subset of the students in grades 4 through 8 of the 1988 Fall and Spring standardization samples of the Comprehensive Tests of Basic Skills, Fourth Edition (CTBS/4, CTB/McGraw-Hill, 1989). Performance on the following five tests was considered: Reading Comprehension, Spelling, Language Expression, Mathematics Concepts and Applications, and Science. The test levels were 14 (grade 4), 15 (grade 5), 16 (grade 6), 17/18 (grades 7 and 8). The tests administered to these students are called Benchmark tests and each has 40 items. Each Benchmark test is the combination of two 20-item Survey tests, called Forms A and B. These forms were constructed as part of a 1987 national tryout study to be as parallel as possible in terms of content coverage, difficulty, and standard error of measurement functions. The items from Survey Forms A and B are intermingled to create the Benchmark.

All items had four choices. Instructions in the CTBS/4 Examiner's Manual (CTB/McGraw-Hill, 1988) related to guessing were as follows:

> Guessing. Encourage students to attempt all items, even those dealing with content that is unfamiliar. Students should be encouraged to make their most careful choice for each item and then go on. Avoid using the term guessing as this may encourage random marking.
>
> Random Marking. Students should not mark answers randomly. Random marking indicates that the student has no understanding of the item. If you determine that a student has marked answers randomly, treat the test as a nonvalid test. (p. 4)

There were from 12,787 to 15,764 students in each of the standardization samples at the 10 testing times (Fall and Spring at each of 5 grades). Random samples of 1000 students from each of the 10 testing times were chosen. In the present study, a student was included in the analyses for a test only if he/she had a valid attempt at both Survey forms; a valid attempt is defined as answering at least 5 items or getting at least 1 item correct. (The same type of screening occurs in the norming and customer scoring of these tests.) This screening produced samples of 959 to 994 for Reading Comprehension, Spelling, Language Expression, and Math Concepts and Applications, and from 914 to 991 for Science.

Scores

Each form of each Survey test was scored seven ways. The first was the traditional number-correct score:

NC = number of correctly answered items

In addition each form of each Survey test was also scored the following six ways:

SSip = Scale score based on item-pattern scoring
$\hat{NC}$ip = Expected number-correct based on SSip
GEip = Grade Equivalent score based on SSip

SSnc = Scale score based on number-correct scoring
$\hat{NC}$nc = Expected number-correct based on SSnc
GEnc = Grade Equivalent score based on SSnc

These 6 scores are the combination of 3 metrics and 2 data types:

		Data	
		Number Correct	Item Pattern
Metric	Scale Score	SSnc	SSip
	Number Correct	$\hat{NC}$nc	$\hat{NC}$ip
	Grade Equivalent	GEnc	GEip

Figure 2 may help in understanding the following description of the how these scores were obtained for each 20-item form of the Survey.

Insert Figure 2 about here

SSip (the "item-pattern scale score") is the scale score based on maximum likelihood three-parameter IRT item-pattern scoring (Lord, 1980, pp. 58-59) and it is obtained with a computer algorithm, using the item parameters (a_j, b_j, and c_j, j=1,..., 20) for that form of the Survey. These item parameters had been estimated based on samples of 4000 to 7000 students who participated in the standardization of the tests. The parameters used in this study are the same as those used in the norming and customer scoring of the tests.[3]

SSip was converted to a score in the number correct metric, $\hat{NC}$ip, called the "item-pattern expected number correct score." Such an estimated number correct score is sometimes called an estimated true score (e.g., Lord, 1980, p. 49) and is the sum

(over the items) of the predicted item p-values based on the item parameters and the student's scale score:

$$\hat{NC} = \sum_{j=1}^{20} P_j(SS), \qquad (1)$$

where

$$P_j(SS) = c_j + [1-c_j]/[1+\exp(-1.7a_j(SS-b_j))].$$

To obtain $\hat{NC}ip$, SS = SSip in Equation (1). $\hat{NC}ip$ scores look just like number-correct scores, with two exceptions: They are continuous scores that can take on non-integer values, and they cannot be lower than Σc_j, which will be called the "guessing level."[4] In the examples in Table 2, the student with SSip = 680 has $\hat{NC}ip$ = 11.3, and the student with SSip = 693 has $\hat{NC}ip$ = 12.6.

GEip was obtained by entering the appropriate published SS-to-GE conversion table with the student's item-pattern scale score (SSip).

Starting with the number-correct data, scores in the same three metrics were obtained. SSnc is the "number-correct scale score" obtained from the published NC-to-SS conversion table for that form of the Survey, as in the example in Table 1. These tables produce maximum likelihood scale scores; the tables themselves had been produced as part of the standardization of the tests following the procedure described by Yen (1984).

$\hat{NC}nc$ is an estimated number-correct score obtained from Equation (1) using SS = SSnc. $\hat{NC}nc$ takes on discrete values, and for every student $\hat{NC}nc$ is essentially the same as NC except that $\hat{NC}nc$ never has a value below that expected by guessing.[5] For example, if Σc_j = 4.9, students with NC scores of 0, 1, 2, 3, and 4 have $\hat{NC}nc$ = 4.9; students with NC = 5, have $\hat{NC}nc \doteq 5$; students with NC = 6, have $\hat{NC}nc \doteq 6$, and so forth. The reliability of the $\hat{NC}nc$ scores will be compared with that of the NC scores to see if scores below the "guessing level" contribute to score reliability.

GEnc is obtained by entering the published SS-to-GE table with the student's number-correct scale score (SSnc).

Reliability

In summary, each form of the Survey test was scored seven ways; for each type of score, the Pearson correlation between Form A and Form B was the empirical reliability, $r_{XX'}$, for that score.

Relative Efficiency

In order to provide a description of the magnitude of the difference in measurement accuracy provided by item-pattern and number-correct scoring, relative efficiencies were calculated. The relative efficiency of item-pattern to number-correct scoring is defined as follows:

$$RE(ip \text{ to } nc) = \frac{\hat{\sigma}^2_{E(nc)}}{\hat{\sigma}^2_{E(ip)}}$$

$$= \frac{\hat{\sigma}^2_{X(nc)}[1-r_{XX'(nc)}]}{\hat{\sigma}^2_{X(ip)}[1-r_{XX'(ip)}]},$$

where $\hat{\sigma}^2_X$ is the observed variance of the score (averaged over Forms A and B) and $r_{XX'}$ is the score's reliability. The relative efficiencies were calculated separately for each metric.[6] The relative efficiency gives an estimate of the effect of the difference in measurement error as a proportional increase in the number of items. For example, a RE(ip to nc) of 1.15 means that the use of item-pattern scoring in place of number-correct scoring is equivalent, on the average, to a 15 percent increase in test length.

Results

Background information

Summary descriptive statistics are presented by grade for the standardization samples in the CTBS/4 Technical Bulletin Number 1 (CTB Macmillan/McGraw-Hill, 1989). Those statistics are presented in number correct and scale score units and include means, standard deviations, item p-values, b_i values, and mean c_i values. The summary statistics for the present study were quite similar to those in the Technical Bulletin. Some of these statistics, as well as some data from the norms, are relevant to the present analyses and can be summarized for the tests and grades in this study.

The standardization mean proportion correct scores ranged from .46 to .66 in the Fall and from .51 to .70 in the Spring. The average c_i values ranged from .20 to .23.

These tests showed minimal floor or ceiling effects; that is, in the normative distributions scores of $\Sigma c_i + 1$ corresponded to no higher than the 6-th percentile in the Fall and the 4-th percentile in the spring; number-correct scores of 1 less than perfect corresponded to the 99-th percentile.

These tests had minimal speededness effects. The standardization completion rates (percent of students answering the last item) ranged from 93 to 99, with a median of 97.

Reliability

Tables 3 to 7 contain the empirical reliabilities for each of the five content

---

Insert Tables 3 to 7 about here

---

areas. At the bottom of each table are averages of these reliabilities. (While reliabilities are presented here to two decimal places, in making comparisons or calculating statistics, at least three decimal places were used.) Several results are evident from these reliabilities:

1) In the vast majority of cases the scores in the number correct metric (NC, $\hat{N}$Cnc, $\hat{N}$Cip) had higher Pearson correlations than scores in the scale score or grade equivalent metrics. The scores in the grade equivalent metric usually had higher correlations than scores in the scale score metric, although in 33 percent of the cases, the scale scores had higher correlations than the corresponding grade equivalents.

It should be kept in mind that for a given type of data, the three metrics rank order students in the same way and must have the same reliability in terms of rank ordering students. Rank order correlations are discussed further later in this paper.

2) $\hat{N}$Cnc scores were more reliable than NC scores in 43 out of 50 replications. Recall than $\hat{N}$Cnc scores are essentially the same as the NC scores, with the exception that $\hat{N}$Cnc scores do not go below the guessing level. Thus, there appears rarely to be reliable information about student performance provided by scores below the guessing level.

3) Scoring that used item-pattern data was clearly more reliable than number-correct scoring:

a) $\hat{N}$Cip scores were more reliable than $\hat{N}$Cnc scores in 50 out of 50 replications (50 replications = 5 content areas by 10 testing times.)

b) $\hat{N}$Cip scores were more reliable than NC scores in 50 out of 50 replications.

c) SSip scores were more reliable than SSnc scores in 49 out of 50 replications.

d) GEip scores were more reliable than GEnc scores in 49 out of 50 replications.

Table 8 displays the mean reliabilities, by content area, with the means taken over the ten testing times. These means show the same trends as those just described.

Insert Table 8 about here

Relative Efficiency

The relative efficiency of item-pattern to number-correct scoring was calculated for each of the 5 content areas for each of the 10 testing times. These relative efficiencies are presented in Figure 3; again it is clear that item-pattern scoring is more efficient than number-correct scoring in every metric. On the average, item-pattern scoring produced an increase in accuracy equivalent to 18 percent more items in the number-correct metric, 20 percent more items in the scale score metric, and 15 percent more items in the grade equivalent metric. Increases in efficiency due to item-pattern scoring tended to be greater in the Fall than in the Spring.

Insert Figure 3 about here

Additional Metrics

Pearson reliabilities were obtained in two additional metrics for the 25 Spring replications: percentile and NCE. These metrics had lower Pearson reliabilities than the number-correct metric. Pearson reliabilities in the percentile metric were lower than those in the number-correct metric in 20 of 25 replications for item-pattern scoring (PERCENTILEip vs $\hat{N}$Cip) and in 18 of 25 for number-correct scoring (PERCENTILEnc vs $\hat{N}$Cnc). Pearson correlations in the NCE metric were lower than those in the number-correct metric for 24 of 25 replications for both item-pattern scoring (NCEip vs $\hat{N}$Cip) and number-correct scoring (NCEnc vs $\hat{N}$Cnc).

Item-pattern scoring was more reliable than number-correct scoring in 25 of 25 replications in both the percentile (PERCENTILEip vs PERCENTILEnc) and NCE (NCEip vs NCEnc) metrics. The increase in reliability due to item-pattern scoring was equivalent, on the average, to a 15 percent increase in test length in the percentile metric and 15 percent in the NCE metric.

Rank Order Correlations

In addition to the Pearson correlations reported in Tables 3 to 8, Spearman rank order correlations were calculated between forms. Within the item-pattern scoring method, the five metrics ($\hat{N}$Cip, SSip, GEip, NCEip, and PERCENTILEip) rank order students in the same way[2] and have the same reliability in terms of rank ordering students, which will be called RANKip. Similarly, using number-correct data, the five metrics ($\hat{N}$Cnc, SSnc, GEnc, NCEnc, and PERCENTILEnc) must have the same rank order correlation, RANKnc.

For each of the 50 tests and testing times RANKip and RANKnc were obtained. The rank order correlations were similar in magnitude to the Pearson correlations in the number-correct metric. The rank order correlation using item-pattern data (RANKip) was higher than the rank order correlation using number-correct data (RANKnc) in 50 out of 50 replications.

Generality of the Results

Previous Research

The increases in relative efficiency for item-pattern scoring tended to be greater in the Fall than the Spring; this result is expected because there is a greater theoretical advantage in item-pattern scoring among low scoring students than high scoring students (Yen, 1984). In rough terms three-parameter item-pattern scoring "takes guessing into account," so that it theoretically has greater effect among lower scoring students.

Some people have the misconception that because item-pattern scoring "takes guessing into account" it lowers students' scores relative to number-correct scoring. Theoretical and empirical investigations have made it very clear that item-pattern and number-correct scoring are tau-equivalent; that is, the two scoring procedures produce the same scores on the average at every scoring level. This average equivalence is found in representative national samples (Yen, 1984), samples of Black students (Green & Yen, 1983), and in every set of CTB customer data that has been examined. Analyses of the present data sets (not reported here) found tau equivalence in every metric comparable to that previously reported.

The increases in relative efficiency for item-pattern over number-correct scoring tend to be greater in the present study than in Yen (1984), where the relative efficiencies ranged from 1.00 to 1.13. The tests in the present study tend to be more difficult and have a greater variation in item difficulty than those in Yen (1984). There is theoretical basis for both these differences contributing to greater relative efficiency. Item-pattern scoring has theoretically greater effect among students for whom the test is more difficult; this effect can be seen in the standard error functions pictured in Yen (1984). Also, the more that items vary in their statistical characteristics, including difficulty, the more that differences in response patterns will matter (Lord, 1980, pp. 58-60).

The finding that the number-correct metric is more reliable than the scale score metric is consistent with previously reported comparisons based on item-pattern scoring (Lord, 1983b; Yen, 1984, 1987) and number-correct scoring (Yen, 1984).

Completion of tests

The items in the two Survey forms were intermingled in the Benchmark. Thus, the scores on the two Survey forms are not differentially affected by testing occasion, as they would be if they were administered separately. These tests have minimal speededness, and given the data screening that was conducted for this study, artifactual inflation of the reliabilities due to non-completion of the tests was minimal and affected number-correct scoring and item-pattern scoring in a similar way. More extensive data screening was conducted on a random subset of the replications, but it did not affect the conclusions of the study.

Parameter Estimation

The students in the present study were part of the standardization samples used to estimate the item parameters. The question therefore arises of whether the gains in efficiency obtained from item-pattern scoring would be seen in other samples of students who were not included in the parameter estimation. In partial but not complete response to this question it can be noted that the 1000 students in each replication in this study were a small part of the parameter estimation samples, which ranged from 4000 to 7000. Also, the parameter estimation pooled data over at least two grades, over Fall and Spring, as well as over on-level and inter-level linkage tests (e.g., level 14, level 13/14 linkage, level 14/15 linkage). (For details see CTBS/4 Technical Bulletin 1, p. 32.) It is noteworthy that even though the item parameters were obtained by pooling data from different grades and tests, they were sufficiently generalizable that their use produced the substantial increases in efficiency seen with item-pattern scoring.

Conclusions

The Rasch Model

Item-pattern scoring has a theoretical advantage in score accuracy for the two- and three-parameter IRT models but not for the one-parameter (Rasch) model. For the Rasch model, it is assumed that all items have the same discriminations and there is no chance of very low scoring students getting items correct through guessing. When the Rasch model is true, the NC score contains all relevant score information and item-pattern scoring produces no increases in score accuracy. Lord (1983a) points out that even if a two- or three-parameter is more appropriate than the Rasch model, if the additional item parameters are not well estimated, say because small sample sizes are used (e.g., $N < 200$ for the two-parameter model), then the theoretical advantages of item-pattern scoring may not be gained in practice.

It is clear from the present study that the additional parameters were sufficiently well estimated that their use in item-pattern scoring produced substantial empirical gains in score accuracy. All the hundreds of multiple-choice tests that compose CTBS/4 contain items that clearly have non-zero guessing levels and a variety of discriminations. (The same is true for the hundreds of tests in CTBS, Forms U and V, and the California Achievement Tests, Forms E and F.) Thus, the Rasch model is not suitable for these tests. However, if the Rasch model had been applied to these tests, it would have produced the amount of score reliability seen in number-correct scoring. Thus, the use of the three-parameter item-pattern scoring produced increases in score reliability over that would be provided by the Rasch model comparable to an average 15 to 20 percent increase in test length.

Scores Below the "Guessing Level"

Recall that NC and $\hat{NC}nc$ scores were essentially the same, except that the $\hat{NC}nc$ scores could not go below Σc_i. Was important information lost by removing differences among students below the "guessing level"? In 43 out of 50 replications, the answer was, "No"-- $\hat{NC}nc$ scores were actually more reliable than NC scores. In the other 7 replications the NC and $\hat{NC}nc$ reliabilities were almost indistinguishable.

Thus, scores below the guessing level usually had above average amounts of error variance, and the variation among them usually had a negative impact on over-all reliability.

Choosing a Metric

If maximizing the Pearson reliability is the criterion by which a metric is to be chosen, the number-correct metric is to be preferred over the grade equivalent, percentile, scale score, and NCE metrics. However, in most cases maximizing Pearson reliability is not the most important criterion for choosing a metric. It would not make sense for educational measurement to abandon all use of percentiles, grade equivalents, scale scores, and normal curve equivalents because the number-correct metric has higher Pearson correlations. Number-correct scores cannot be meaningfully combined or compared across tests of different difficulty levels, and number-correct scores carry no normative information. The normative information provided by percentiles, GEs, and NCEs is important and useful.

The continuity provided by scale scores is practically and conceptually important for relating results from different test levels and providing an cohesive framework for scoring and analysis. For example, as described in Figure 2 the scale score provides the vehicle for taking advantage of the information in the item response patterns. Score reports usually contain scores in several metrics so that users can choose the metric that is most useful for a particular decision. Most practitioners do not find scale scores to be particularly meaningful, and they commonly pay closest attention to percentiles and GEs. It can be noted that the National Assessment of Educational Progress reports scale scores that are linearly related to the IRT scale score;[7] as interpretive information for this scale becomes more widely disseminated, it is likely that these scores will come to have more meaning for a wider audience.

Choosing a Scoring Method

It is clear that from the point of view of maximizing Pearson and rank order reliability, item-pattern scoring is to be preferred over number-correct scoring in every metric examined. In other words, regardless of the metric in which scores are to be reported or analyzed, scores in that metric are expected to be more reliable if they are reached via item-pattern scoring than number-correct scoring. The increases in reliability produced by item-pattern scoring for the present tests were equivalent, on the average, to that expected by a 15 to 20 percent increase in test length.

Technical Notes

[1]An alternative description of the difference between number-correct and item-pattern scoring is that they are based on the same data or information, but they differ in terms of how much of this information they use. That is, they both start off with the correct or incorrect response to each of a set of items, the identity of each of those items, and the parameters for each of those items. Number-correct scoring discards the information contained in which items were correct and which were incorrect.

[2]Except for ties.

[3]A maximum likelihood procedure cannot produce finite scale scores for students with perfect scores or scores below the "guessing level." Also, while maximum likelihood scale scores are available for students with extreme scores other than zero or perfect scores, these estimates can have standard errors that are very large and differences between the these extreme values have little meaning. Therefore, as part of the standardization, a rational but necessarily non-maximum likelihood process was followed to set the lowest obtainable scale score (LOSS) and highest obtainable scale score (HOSS) for each level of each subtest; for a given level, the same LOSS and HOSS values are used for all forms and both number-correct and item-pattern scoring. These values are the same ones used in the customer scoring of these tests. For example, in Tables 1 and 2 the LOSS and HOSS are 528 and 836, respectively.

[4]What has been called here the "guessing level" is the lower asymptote of the test characteristic function and is the number correct score expected for students with very low scale scores. This guessing is not assumed to be random. In fact, the c_i values tend to be lower than what would be expected from random guessing; for example, for four-choice items the c_i values tend to be less than 1/4. These values reflect the fact that wrong answers are written to correspond to important errors.

[5]The relationship between the estimated true score based on SSnc and the observed number-correct score is detailed in Yen (1984, p. 109).

[6]Lord (1980, Chapter 6) demonstrates that theoretical relative efficiencies, expressed as a function of score level, are metric independent. In this study the relative efficiencies were expected to vary somewhat over metric because they were empirical rather than theoretical. Also, the values used in this paper were ratios of average error variances (with each average taken over the given sample of students) rather than averages (taken over students) of ratios of error variances.

[7]The NAEP scale is a linear transformation of an estimated true score on a hypothetical set of items; these hypothetical items are specified in such a way that the NAEP scale is linearly related to the IRT scale score within the range of observed scores.

References

CTB/McGraw-Hill. (1988). Comprehensive Tests of Basic Skills, Fourth Edition, Standardization Examiner's Manual, Levels F-L. Monterey, CA: Author.

CTB/McGraw-Hill. (1989). Comprehensive Tests of Basic Skills, Fourth Edition. Monterey, CA: Author.

CTB Macmillan/McGraw-Hill. (1989). Comprehensive Tests of Basic Skills, Fourth Edition, Technical Bulletin 1. Monterey, CA: Author.

Green, D.R., & Yen, W.M. (1983). Number correct versus pattern scoring: Results for ethnic groups. Paper presented at the annual meeting of the National Council on Measurement in Education.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F.M. (1983a). Small N justifies Rasch model. In D.J. Weiss (Ed.), New horizons in testing (pp. 51-62). New York: Academic Press.

Lord, F.M. (1983b). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. Psychometrika, 48, 233-245.

Qualls-Payne, A., & Hoover, H. D. (1987). A comparison between test reliabilities for number-correct and IRT item-pattern scoring. Paper presented at the annual meeting of the National Council on Measurement in Education.

Yen, W.M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. Journal of Educational Measurement, 21, 93-111.

Yen, W.M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. Psychometrika, 52, 275-291.

Table 1

Example of Using a Table to Convert From One Metric to Another Using the Number-Correct Scoring Method

NC	SS	Percentile	GE	NCE
0-4	528	1	1.4	1
5	556	2	1.6	9
6	610	7	2.2	19
7	633	11	2.5	24
8	648	15	2.7	28
9	659	19	2.9	32
10	669	23	3.2	34
11	678	28	3.6	38
12	687	33	3.9	41
13	697	40	4.3	45
14	707	49	4.7	49
15	717	57	5.2	54
16	729	69	5.9	60
17	741	79	6.8	67
18	756	89	8.8	76
19	779	97	11.2	89
20	836	99	12.9+	99

Table 2

Example of Expanding a Conversion Table to Include Item-Pattern Scores

SS	Percentile	GE	NCE	$\hat{NC}^*$
528	1	1.4	1	4.9
556	2	1.6	9	5.0
610	7	2.2	19	6.1
633	11	2.5	24	7.1
648	15	2.7	28	8.1
659	19	2.9	32	9.0
669	23	3.2	34	10.1
678	28	3.6	38	11.0
680	29	3.7	38	11.3
687	33	3.9	41	12.0
693	37	4.2	43	12.6
697	40	4.3	45	13.1
707	49	4.7	49	14.1
717	57	5.2	54	15.0
729	69	5.9	60	16.0
741	79	6.8	67	17.0
756	89	8.8	76	18.0
779	97	11.2	89	19.0
836	99	12.9+	99	19.8

Note. In the actual item-pattern table, hundreds of scale scores from 528 to 836 would be included.

*Expected number-correct given the scale score.

Table 3

Survey Reliabilities: Reading Comprehension

		Score						
Level	Grade	NC	$\hat{N}$Cnc	$\hat{N}$Cip	SSnc	SSip	GEnc	GEip
14	4.2	.82	.83	.84	.75	.79	.73	.73
	4.8	.82	.82	.84	.75	.76	.71	.73
15	5.2	.78	.78	.82	.70	.74	.75	.78
	5.8	.79	.79	.82	.72	.76	.75	.77
16	6.2	.75	.75	.80	.64	.69	.73	.77
	6.8	.77	.78	.81	.65	.71	.77	.80
17/18	7.2	.77	.77	.79	.66	.70	.77	.78
	7.8	.77	.78	.80	.67	.71	.78	.80
	8.2	.79	.80	.82	.69	.73	.78	.80
	8.8	.81	.81	.82	.73	.73	.80	.82
	Mean:	.79	.79	.82	.70	.73	.76	.78

Table 4

Survey Reliabilities: Spelling

		Score						
Level	Grade	NC	$\hat{N}$Cnc	$\hat{N}$Cip	SSnc	SSip	GEnc	GEip
14	4.2	.75	.75	.80	.75	.80	.72	.75
	4.8	.77	.77	.82	.76	.82	.74	.77
15	5.2	.76	.76	.80	.72	.78	.76	.79
	5.8	.77	.77	.81	.74	.80	.74	.77
16	6.2	.71	.71	.77	.70	.77	.70	.76
	6.8	.70	.70	.76	.70	.77	.69	.73
17/18	7.2	.73	.73	.77	.71	.75	.71	.75
	7.8	.72	.71	.77	.72	.78	.70	.76
	8.2	.72	.72	.78	.71	.78	.69	.75
	8.8	.77	.77	.80	.76	.80	.74	.78
	Mean:	.74	.74	.79	.73	.79	.72	.76

Table 5

Survey Reliabilities: Language Expression

Level	Grade	Score						
		NC	$\hat{NC}$nc	$\hat{NC}$ip	SSnc	SSip	GEnc	GEip
14	4.2	.86	.87	.88	.80	.82	.80	.82
	4.8	.87	.88	.88	.80	.82	.80	.82
15	5.2	.83	.83	.85	.79	.81	.79	.80
	5.8	.82	.83	.85	.79	.83	.81	.82
16	6.2	.78	.78	.82	.74	.80	.76	.79
	6.8	.75	.75	.79	.73	.78	.74	.78
17/18	7.2	.75	.75	.82	.72	.79	.74	.80
	7.8	.74	.74	.81	.73	.79	.71	.79
	8.2	.75	.75	.81	.73	.79	.73	.80
	8.8	.79	.79	.83	.78	.81	.76	.81
	Mean:	.79	.80	.84	.76	.80	.76	.80

Table 6

Survey Reliabilities: Mathematics Concepts & Applications

Level	Grade	Score						
		NC	$\hat{NC}$nc	$\hat{NC}$ip	SSnc	SSip	GEnc	GEip
14	4.2	.71	.72	.78	.65	.73	.72	.76
	4.8	.74	.74	.77	.69	.74	.72	.71
15	5.2	.75	.75	.79	.68	.73	.74	.77
	5.8	.77	.77	.79	.71	.73	.76	.78
16	6.2	.79	.80	.82	.73	.75	.79	.81
	6.8	.77	.77	.79	.70	.71	.76	.77
17/18	7.2	.74	.75	.80	.65	.70	.74	.79
	7.8	.79	.79	.82	.72	.75	.78	.80
	8.2	.80	.80	.83	.72	.76	.79	.81
	8.8	.82	.83	.86	.74	.79	.82	.85
	Mean:	.77	.77	.80	.70	.74	.76	.78

Table 7

Survey Reliabilities: Science

		Score						
Level	Grade	NC	$\hat{NC}$nc	$\hat{NC}$ip	SSnc	SSip	GEnc	GEip
14	4.2	.74	.74	.77	.70	.73	.71	.74
	4.8	.78	.78	.81	.72	.76	.76	.78
15	5.2	.68	.68	.74	.61	.68	.68	.74
	5.8	.69	.70	.74	.62	.67	.69	.74
16	6.2	.73	.73	.78	.65	.71	.73	.77
	6.8	.69	.70	.74	.62	.67	.69	.73
17/18	7.2	.69	.70	.73	.61	.63	.68	.70
	7.8	.73	.73	.75	.67	.71	.71	.74
	8.2	.73	.74	.76	.68	.71	.72	.74
	8.8	.75	.76	.77	.72	.73	.74	.75
	Mean:	.72	.73	.76	.66	.70	.71	.74

Table 8

Survey Reliability

Means Taken Over Ten Testing Times

Reading Comprehension

		Data		
		Number Correct	Item Pattern	Mean
Metric	SS	.70	.73	.72
	NC	.79	.82	.80
	GE	.76	.78	.77
Mean		.75	.78	

Spelling

		Data		
		Number Correct	Item Pattern	Mean
SS		.73	.79	.76
NC		.74	.79	.76
GE		.72	.76	.74
Mean		.73	.78	

Language Expression

		Data		
		Number Correct	Item Pattern	Mean
Metric	SS	.76	.80	.78
	NC	.80	.84	.82
	GE	.76	.80	.78
Mean		.77	.81	

Mathematics Con. & Ap.

		Data		
		Number Correct	Item Pattern	Mean
SS		.70	.74	.72
NC		.77	.80	.79
GE		.76	.78	.77
Mean		.74	.78	

Science

		Data		
		Number Correct	Item Pattern	Mean
Metric	SS	.66	.70	.68
	NC	.73	.76	.74
	GE	.71	.74	.73
Mean		.70	.73	

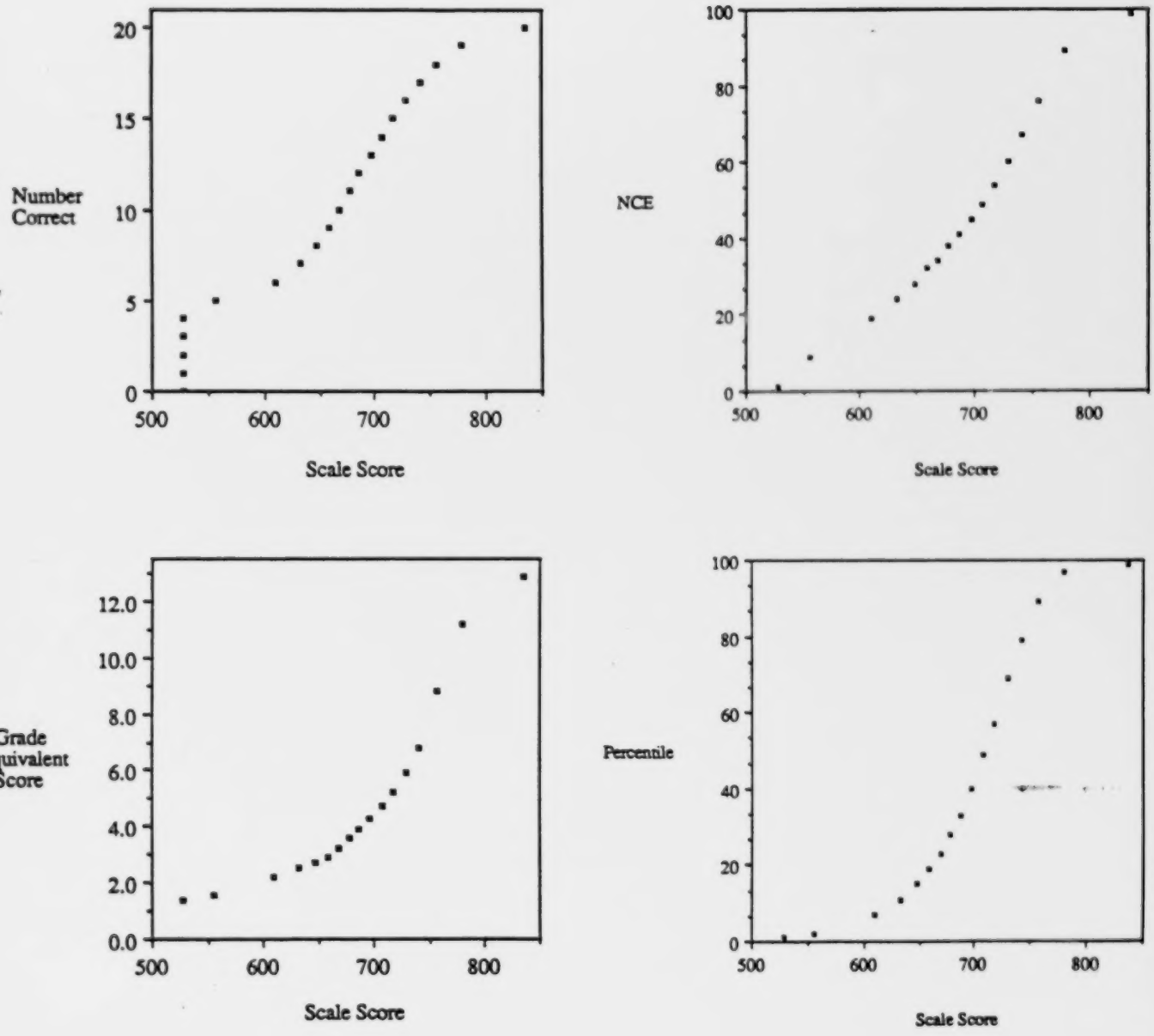

Figure 1. Examples of relationships between metrics

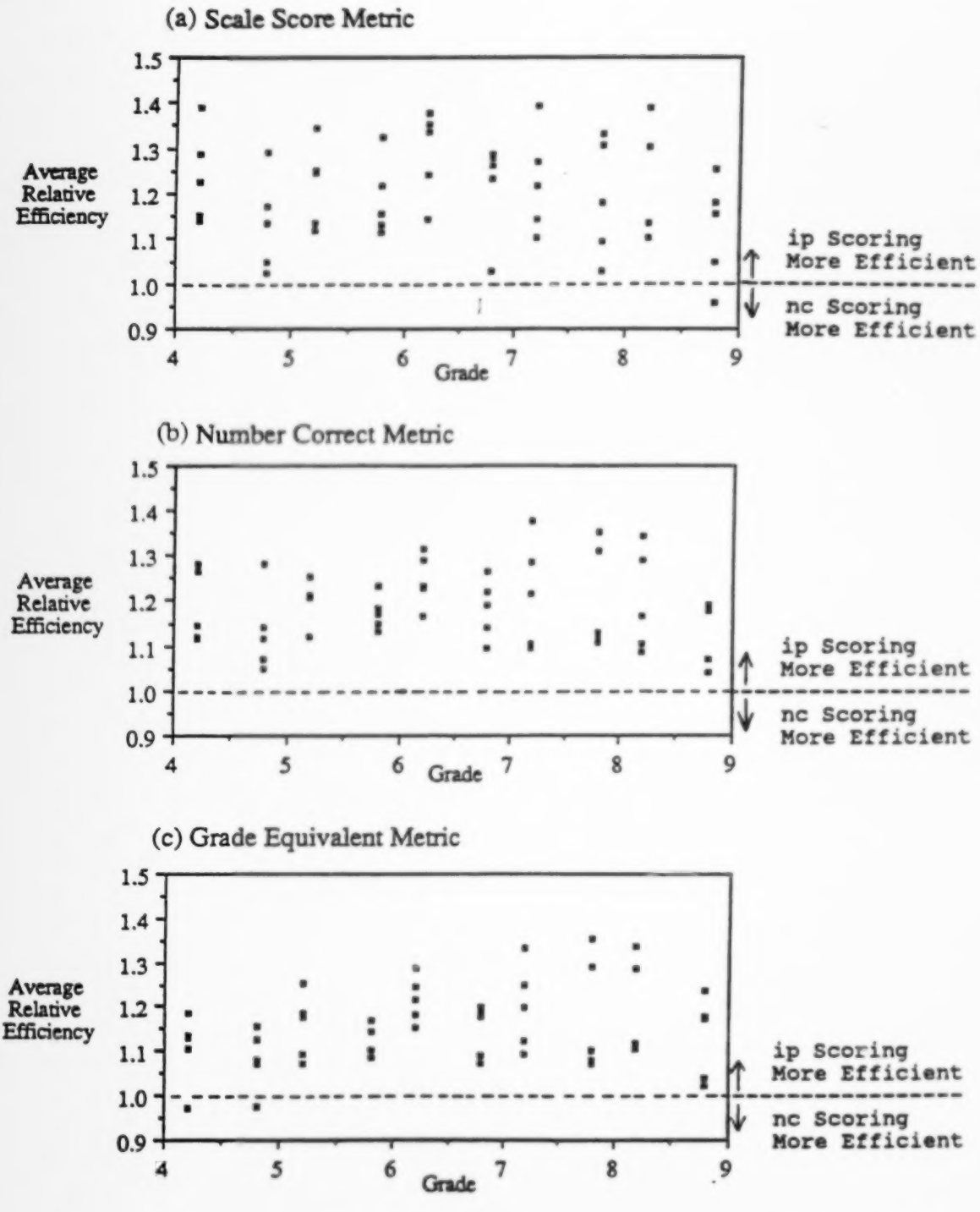

Figure 3. Average relative efficiencies of item-pattern (ip) to number-correct (nc) scoring.

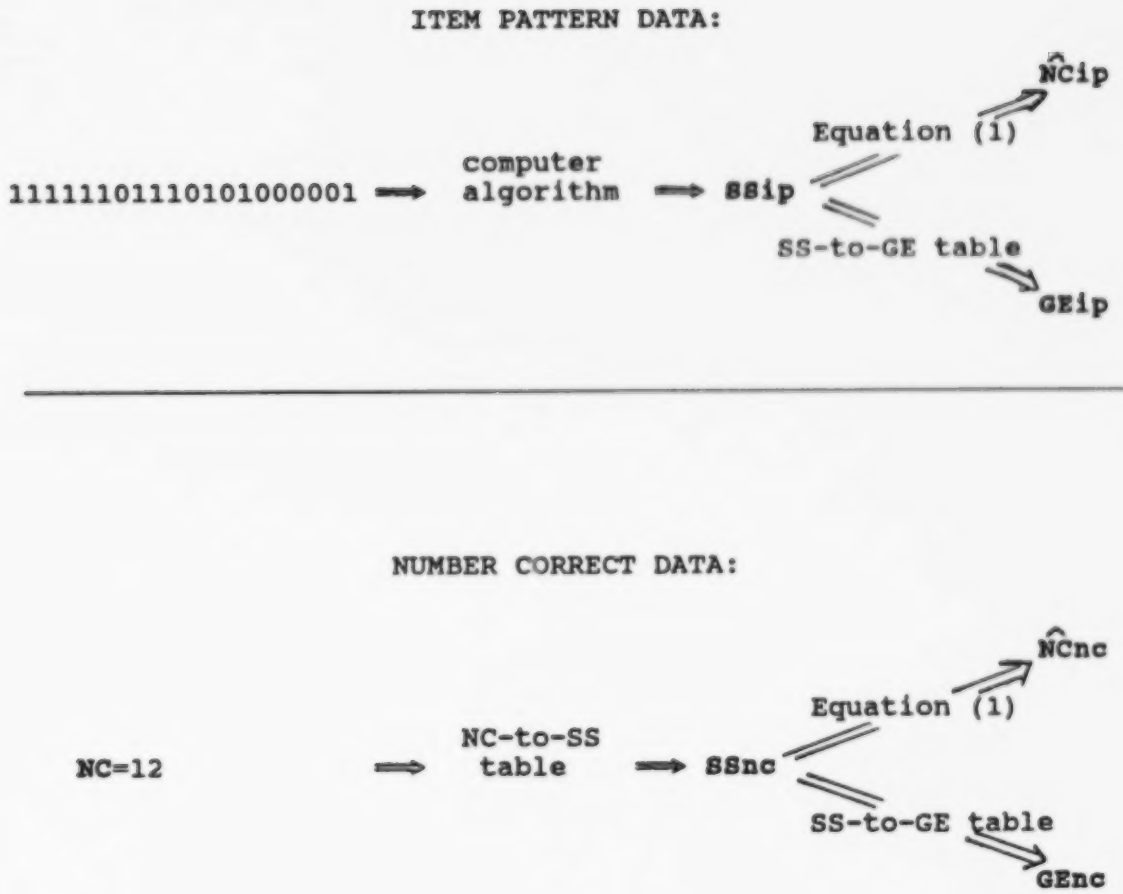

Figure 2. Procedure for obtaining scores for this study.